DOCUMENT RESUME

ED 452 205 TM 032 490

AUTHOR De Champlain, Andre F.; Gessaroli, Marc E.; Floreck, Lisa M.
TITLE Assessing the Impact of Standardized Patient Variability on Examination Mastery-Level Decision Consistency Rates.
PUB DATE 2000-04-25
NOTE 18p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Classification; *Interrater Reliability; *Licensing Examinations (Professions); Medical Education; *Physicians; Simulation
IDENTIFIERS Case Analysis; *Clinical Competence; *Standardized Patients

ABSTRACT

The purpose of this study was to estimate the extent to which recording variability among standardized patients (SPs) has an impact on classification consistency with data sets simulated to reflect performances on a large-scale clinical skills examination. SPs are laypersons trained to portray patients in clinical encounters (cases) and to record as well as rate examinee behaviors using case-specific checklists and rating scales. The conditions modeled were intended to approximate those that might occur with SP testing as part of the United States Medical Licensing Examination with populations of U.S. medical school graduates and populations that contain U.S. graduates and international medical graduates. Initially, proficiencies were randomly generated for a n(0,1) distribution for 100,000 simulees. Conditions that were simulated, in terms of the characteristics of percent-correct score distributions, overall failure rate, and amount of variability among recordings were similar to those noted in past research with this type of examination. Results suggest that the impact of recording discrepancies on overall misclassification rates is minimal for a (homogeneous) population resembling first time U.S. medical school graduates. Findings differed for the more heterogeneous population, and the total misclassification rate was 2.6% higher in the baseline condition. Results suggest that errors of commission similar to those simulated in this investigation could have serious consequences on the false positive rate for a more heterogeneous group of examinees unless some adjustments are made to account for SP rater variability. (Contains 2 tables and 30 references.) (SLD)

Assessing the Impact of Standardized Patient Variability on Examination Mastery-Level Decision Consistency Rates

André F. De Champlain, Marc E. Gessaroli, & Lisa M. Floreck

National Board of Medical Examiners

Paper presented at the meeting of the National Council on Measurement in Education

New Orleans, LA

Tuesday, April 25, 2000

RUNNING HEAD: Impact of SP Variability on Decision Consistency Rates

Assessing the Impact of Standardized Patient Variability on Examination Mastery-Level Decision Consistency Rates

Standardized patient (SP) examinations are being used with increasing frequency by medical schools and testing organizations to assess the clinical skills of medical students in a range of simulated doctor-patient encounters (AAMC, 1998). SPs are laypersons trained to portray patients in clinical encounters (referred to as cases) and to record as well as rate examinee behaviors using case-specific checklists and rating scales (Barrows, 1987).

Although a large body of research has been dedicated to assessing the reliability of SP examination scores and related inferences, relatively few studies have addressed some more basic issues critical for all clinical skills assessments that utilize SPs as recorders of student behavior (Swanson & Norcini, 1989; Swanson, Norman, & Linn, 1995; Vu & Barrows, 1994). One of these issues is the extent to which SP checklist recording and rating discrepancies impact upon pass/fail decisions. This issue is of particular relevance for testing programs that train several SPs for each case where it must be shown that the likelihood of passing a case and/or the examination is unrelated to the particular cohort of patients that a student might have seen during the test. Although precautions can be adopted to minimize this risk (e.g., randomly assigning students to different SPs for each case and excluding problematic SPs), discrepancies in recording and rating accuracy could still have deleterious effects on the probability that a given candidate will pass the examination.

Investigations that estimated checklist item-level recording accuracy by comparing SP responses to those of a scoring key have generally reported high proportion of agreement rates, ranging from the mid .70s to the upper .90s (De Champlain, Margolis, King, & Klass, 1997; Vu, Marcy, Colliver, Verhulst, Travis, Barrows, 1992). Findings reported in past generalizability studies also seem to suggest that there is little overall checklist score variability attributable to SPs (Swanson & Norcini, 1989; Swanson, Norman, & Linn, 1995; van der Vleuten & Swanson, 1990). However, it is important to point out that the small variance components estimated for the *Raters* facet (i.e. SPs) in these studies are not necessarily indicative of perfect agreement among SPs. This is especially

true in light of the inordinately large variance component typically associated with differences in performance from case to case throughout an examination (which is indicative of the content-specific nature of performances; Linn & Burton, 1994). Despite these small variance components, it remains possible that SP recording discrepancies are important with respect to the consistency with which both scores can be rank-ordered and mastery-level decisions can be ascribed.

Researchers who assessed the impact of using multiple SPs per case on the reliability estimates of overall case and component scores (i.e. checklist, written post-encounter note, etc.) have generally reported only modest effects on generalizability coefficient values (Colliver, Marcy, Vu, Steward & Robbs, 1994; Colliver, Morrison, Markwell, Verhulst, Steward, Dawson-Saunders & Barrows, 1990). Similarly, negligible differences were noted in two studies that compared station pass/fail rates across multiple SPs portraying identical cases (Colliver, Robbs, & Vu, 1991; Reznick, Smee, Rothman, Chalmers, Swanson, Dufresne, Lacombe, Baumer, Poldre, Levasseur, Cohen, Mendez, Patey, Boudreau & Bérard, 1992). However, a study undertaken by De Champlain, Macmillan, Klass & Margolis (1999) which looked at the effect of not only using multiple SPs per case but also administration at multiple test sites reported a significant intra-site variability effect. Although test sites on the whole (i.e. inter-site variability) were comparable with respect to recording consistency, differences noted at the individual SP-level (i.e. intra-site variability) were of serious enough concern to warrant further attention (differences in examinee scores ranged from 5% - 10% as a function of the SP cohort encountered).

In particular, assessing the impact of cross-SP variability constitutes a critical first step prior to (polytomous) calibration, scaling and linking efforts. Much of the research dedicated to linking performance assessments has focused on using extensions of equating methods originally devised for use with multiple-choice items (Baghi, Bent, & Delain, 1995; Baker, 1992; Clauser, Ross, Nungester, & Clyman, 1997; Cope, 1995; Hennings & Hirsch, 1996; Huynh & Ferrara, 1994; Kim & Cohen, 1995; Sykes, Yen, & Ito, 1996; Tzou, 1996). Huynh & Ferrara (1994) found the partial-credit model (Masters, 1982) to be useful for equating performance assessments that were moderately difficult and homogeneous, with respect to content. Tzou (1996) reported that polytomous IRT and linear equating procedures provided comparable results with writing assessment data. The

simulation study conducted by Fitzpatrick and Yen (1999), using a 2-PL partial credit model, also provided useful guidelines to practitioners in terms of sample size, test length and reliability requirements. However, as pointed out by Tate (1999) in a recent simulation study, the use of most of these methods is predicated upon the assumption that judges' level of severity in rating performances is comparable from form-to-form. Although this assumption might be plausible for the scoring of simpler essays with analytic keys, differences in stringency noted with more complex performance-based assessments, such as those routinely found in medical education, make this hypothesis tentative at best. The chief concern, as it pertains to equating within a licensure framework, is to ensure that classification of examinees as masters or nonmasters is accurate, especially for those test takers that are in the vicinity of the cutscore. That is, the equating process must yield scores that reflect underlying abilities of examinees with the smallest amount of estimation error.

The purpose of the present research was therefore to estimate the extent to which recording variability among SPs impacts upon classification consistency with data sets simulated to reflect performances on a large-scale clinical skills examination. Specifically, the conditions modeled were intended to approximate those that might occur with SP testing as part of the United States Medical Licensing Examination (USMLE) with the following two populations:

- United States Medical Graduates (USMGs) only (homogeneous population with respect to clinical skill level);
- A combination of International Medical Graduates (IMGs) and USMGs (heterogeneous population with regard to clinical skill level).

In addition to the latter baseline condition, classification consistency was also ascertained with data sets that were simulated to reflect SP recording variability by respectively adding 5% and 10% to the total examination scores simulated with the above described homogeneous and heterogeneous samples of simulees. The addition of 5% and 10% to simulated expected-correct (EPC) scores was meant to reflect errors of commission ("giving the benefit of the doubt to examinees") which are more common with SP tests than errors of omission (Vu, Marcy, Colliver, Verhulst, Travis, & Barrows, 1992). This precursory research is essential in determining whether SP (or even site)-

related adjustments will be necessary in subsequent calibration, scaling and linking processes.

Methods

Description of the NBME Standardized Patient Testing Program

The National Board of Medical Examiners' (NBME) SP examination is designed to assess the clinical skills of candidates for licensure who are about to enter their first postgraduate year. Examinees rotate through a series of clinical scenarios, or cases, and are evaluated on their ability to handle the case using their history-taking (Hx), physical examination (PE), communication (CM) and interpersonal (IP) skills. The first three skills are assessed using case-specific checklists. Two of the three skills are typically assessed per case. Checklists are composed of no more than 25 dichotomously-scored items which indicate whether or not a student has completed a specific task. Interpersonal skills are assessed using a six-item inventory that is identical across cases and scored on a five-point Likert scale. The checklist and inventory are completed by the SP after each 15 minute examinee-patient encounter. Also, each examinee completes an open-ended case-specific post-encounter note following each encounter with the SP which contains a list of their significant positive and negative findings from the encounter. Percent-correct scores, corresponding to the number of points obtained by the examinee out of the total number of available points, are currently reported for three components (checklist, interpersonal inventory and post-encounter note) of the SP test. Additionally, a composite case percent-correct score, corresponding to the mean of the latter three scores, is provided to examinees. It is important to note that the NBME SP test is currently a large-scale research project being considered for inclusion into the USMLE within the next five years.

Data generation model

Initially, proficiencies were randomly generated from a $N(0,1)$ distribution for 100,000 simulees. These values were treated as the "true" proficiency estimates of simulees and denoted by θ_i. Examinee observed scores were generated by adopting the basic tenet of classical test theory, i.e., any latter observed score can be decomposed into a true score component and an independent measurement term such that

$$x_{ij}=r_j\theta_i+\delta_j\sqrt{1-r_j^2} \quad (1)$$

where

r_j = The correlation between the case and the latent trait (the "discrimination" parameter);

δ_j = a random normal deviate corresponding to the error term associated with examinee i's response to case j.

Then, a logistic transformation was applied to the simulated observed scores to bound them on a [0,1] interval using the following model

$$L(Z)=\frac{e^{(1.7*Z)}}{1+e^{(1.7*Z)}} \tag{2}$$

where

$$Z=d+x_{ij} \tag{3}$$

The latter function is equivalent to the following common IRT model

$$Z=ax_{ij}+d \tag{4}$$

McDonald (1985) has shown that function Z, as outlined in equation (4) can be reparameterized as

$$Z=\frac{r_j x_{ij}+t_j}{\sqrt{1-r_j^2}} \tag{5}$$

where t_j is item j's threshold value (difficulty) which relates to a classical estimate of difficulty (proportion correct or p-value) as follows:

$$t_j=\Phi^{-1}(p_j) \tag{6}$$

where Φ^{-1} corresponds to the inverse normal distribution. Also, based on the well known assumption that the IRT discrimination parameter (a) relates to r_{ij} in the following fashion

$$a_j=\frac{r_j}{\sqrt{1-r_j^2}}, \tag{7}$$

Z in equation (5) can be rewritten as

$$Z=a_j x_{ij}+t_j\sqrt{1+a_j^2} \tag{8}$$

Since a=1 in our function (equation (8)), Z reduces to the following,

$$Z=x_{ij}+t_j*\sqrt{2}. \quad (9)$$

The probability of a correct response by examinee i on case j was then estimated by substituting function Z (equation (9)) into equation (2). The expected percent-correct score (EPC) on the total (12-case) examination was subsequently obtained by calculating the mean probability of a correct response across cases for examinee i (and multiplying by 100).

Two additional EPC scores were estimated for each simulee. The first additional score simply entailed adding 5% to the expected percent-correct score initially simulated (baseline condition). The third EPC score was obtained by adding 10% to the measure simulated in the baseline condition. The latter two scores were intended to reflect those that might be obtained when "moderate" and "extreme" errors of commission are noted between two cohorts of SPs assigned to the same set of cases.

Case parameters and test length

Case difficulty and discrimination values selected for the six data sets were similar to those reported in Gessaroli, Swanson, & De Champlain (1998) and reflect those typically encountered with SP examinations. Mean case difficulty and discrimination parameter values were respectively equal to .68 (i.e. 68%) and .50. These values were used to initially simulate an expected percent-correct score for each of the 12 cases using equation (2). Then, the mean expected percent-correct score (across the 12 cases) was treated as the overall expected percent-correct score in all analyses.

Pass/fail cutoff values

True proficiency cutoff scores were set at θ values of -1.64 and -0.84. These cutoff score values result in respectively failing 5% and 20% of simulees. The first failure rate (5%) is that typically encountered with USMG-like SP examinees whereas the second rate (20%) is characteristic of a more heterogeneous population including both IMGs and USMGs.

EPC pass/fail values were derived by estimating the score corresponding to the 5th and 20th percentile in the distribution of observed scores generated. One EPC cutoff score value was estimated for each of the following two conditions:

- Baseline/homogeneous proficiency condition;
- Baseline/heterogeneous proficiency condition.

Analyses

For each of the following six data sets, false positive, false negative and total misclassification rates were computed:

- Baseline/homogeneous proficiency condition;
- Baseline/heterogeneous proficiency condition;
- Moderate SP rater discrepancies/homogeneous proficiency condition;
- Moderate SP rater discrepancies/heterogeneous proficiency condition;
- Extreme SP rater discrepancies/homogeneous proficiency condition;
- Extreme SP rater discrepancies/heterogeneous proficiency condition.

A false positive decision occurs when a true nonmaster (based on their θ estimate) passes the examination based on their EPC score. Conversely, failing a true master based on their EPC score will result in a false negative decision. For the purposes of this study, the false positive rate corresponds to the number of false positives out of the total number of simulees (100,000). Similarly, the false negative rate corresponds to the number of false negative decisions out of the total number of simulees (100,000).

Results

Descriptive statistics for expected percent-correct baseline condition data set

Mean and standard deviation EPC score values were respectively equal to 67.86% and 14.21%. Cronbach's alpha for the simulated 12-case test scores was equal to 0.78 which is typical for scores derived from this type of performance assessment. These values were reported with several past NBME SP prototype data sets (Klass, De Champlain, Fletcher, King, & Macmillan, 1998).

Classification consistency results

False positive, false negative and total misclassification rates for data sets simulated to reflect a homogeneous (USMG-like) population for the baseline, moderate, and extreme rater discrepancy conditions are outlined in Table 1. Overall misclassification rates ranged from 3.7% (moderate rater discrepancy condition) to 4.1% (baseline & extreme rater discrepancy conditions). False positive rates ranged from 1.8% (baseline condition) to 3.9% (extreme rater discrepancy condition) whereas false negative rates varied from 0.2% (extreme rater discrepancy condition) to 2.3% (baseline condition).

Table 2 provides false positive, false negative and total misclassification rates for data sets generated to reflect a more heterogeneous (USMG/IMG-like) population with regard to clinical skill level for the baseline, moderate, and extreme rater discrepancy conditions. Total misclassification rates varied from 10.4% (moderate rater discrepancy condition) to 13.0% (baseline condition). False positive rates ranged from 2.8% (baseline condition) to 10.0% (extreme rater discrepancy condition) whereas false negative rates varied from 1.5% (extreme rater discrepancy condition) to 10.2% (baseline condition).

Discussion

Performance assessments have been incorporated into local as well as large-scale testing programs with increasing frequency over the past decade. Alternative assessments are potentially promising in that certain proficiencies that are difficult (if not impossible) to measure via conventional means might be more readily targeted. Nonetheless, the psychometric properties of performance assessment scores still clearly need to be evaluated to ensure that measures reported to students are accurate representations of their proficiency level on the trait of interest and to preclude ill-informed inferences based on those test scores. The issue of rater comparability is especially critical within the realm of more complex performance based assessments traditionally found in medical education, such as standardized patient examinations. Recording and rating variability across SPs portraying the same clinical scenario is an issue that is of central concern for this type of examination given the implications for calibration, scaling, equating and score validity, more generally.

The purpose of this study was therefore to estimate the extent to which recording variability among SPs impacted upon classification consistency with data sets simulated to reflect performances on a large-scale clinical skills examination. Conditions that were simulated, in terms of the characteristics of percent-correct score distributions, overall failure rate, and amount of variability among recordings were similar to those noted in past research with this type of examination.

Results suggest that the impact of recording discrepancies on overall misclassification rates is minimal for a (homogeneous) population resembling first-time USMGs. Misclassification rates varied by 0.4%, at most. These results are largely attributable to the fact that the cut-score was set at a value that was considerably lower that the mean ability level of this group. Consequently, a very large inter-SP rater variability effect would be needed to yield a substantial shift in decision consistent rates (exceeding the 10% shift instituted in this study). As expected, false positive rates increased as errors of commission became more severe whereas the addition of 10% to the simulated EPC scores resulted in a virtually nil false negative rate.

Findings differed, however, for the more heterogeneous population of simulees where the total misclassification rate was 2.6% higher in the baseline condition. This result was anticipated as adding 10% shifts the EPC score distribution by approximately one standard deviation. It is important to also point out that false positive rates increased significantly as errors of commission became more severe (by more than 7% across rater discrepancy conditions). Given the purpose of this type of examination (medical licensure), the classification error that is most important to minimize is the false positive rate. This is consistent with favoring a policy that would be aimed at protecting the public from (potentially) incompetent physicians. The results outlined in this study suggest that errors of commission similar to those simulated in the current investigation could have dire consequences on the false positive rate for a more heterogeneous group of examinees unless certain adjustments are instituted to account for SP rater variability.

In terms of calibration, scaling and equating tasks, these findings suggest that the impact of errors of commission would probably be modest for a USMG-like population. However, treating each case as identical for calibration, scaling and equating purposes, irrespective of the specific SP portraying the clinical scenario, is perhaps

ill-advised for a more heterogeneous population given the divergences noted in decision consistency across rater discrepancy conditions.

Minimizing intra-site variability should be an important concern for all examinations that use multiple SPs per case so as to ensure that scores reported to students accurately reflect their true clinical skill level. Several methods can be adopted to reduce intra-site SP variability. From a training perspective, a periodic review of videotaped encounters would seem advisable to ensure that SPs are maintaining a high level of recording accuracy across extended periods of time. Additionally, assigning an alternate SP to monitor all encounters might also significantly reduce intra-site variability by including a supervisory element in the process. Finally, deriving a checklist score that is based on the consensus reached by a pair of SPs as to what constituted the actions undertaken by a student in a given encounter might also prove to be a worthwhile strategy to increase the reliability of scores.

Treating each SP-case interaction as a unique and distinct case might also constitute another solution to the problem of SP rater variability in the event that remedial training cannot be offered during the course of the administration. For example, a headache case portrayed by SP#1 might be treated as distinct from the same clinical scenario as depicted by SP #2. This approach is appealing in that the psychometric characteristics (e.g. difficulty and discrimination) of both the case and SP portraying it can be modeled into the calibration and subsequent scaling processes. The disadvantage of implementing such a design is that it can lead to a very sparse data matrix. This is especially true within the context of a national administration where several SPs portray each case at any of up to several dozen test sites. In this instance, the use of a traveling cohort of SPs across administration sites might be worthy of future consideration since they would provide the links needed to undertake concurrent calibrations and subsequent scaling analyses.

Although informative to medical educators involved in SP testing, the findings reported in this study should be interpreted with some degree of caution given the limited nature of the simulations. Future research should be aimed at replicating this research with a wider array of conditions and testing scenarios. Also, the impact of SP variability on parameter estimation needs to be clearly ascertained.

Results of the present research are of great use to the present testing program. It is hoped that these results

will lead to further investigations in an area that is central to not only SP tests but also to all examinations that entail the recording and rating of behavior by human raters.

References

Barrows, H.S. *Simulated (Standardized) patients and other human simulations: A comprehensive guide to their training and use in teaching and evaluation.* (1987). Health Sciences Consortium, Chapel Hill, NC.

Baghi, H., Bent, P., & DeLain M. (1995, April). *A comparison of the results from two equating designs for performance-based student assessments*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Baker, F.B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16*, 87-96.

Clauser, B., Ross, L., Nungester, R., & Clyman, S. (1997). An evaluation of the Rasch model for equating multiple forms of a performance assessment of physicians' patient management skills. *Academic Medicine, 72*, s76-s78.

Colliver, J.A., Robbs, R.S., & Vu, N.V. (1991). Effects of using two or more standardized patients to simulate the same case on case means and case failure rates. *Academic Medicine, 66*, 616-618.

Colliver, J.A., Marcy. M.L., Vu, N.V., Steward, D.E., & Robbs, RS. (1994). Effect of using multiple standardized patients to rate interpersonal and communication skills on intercase reliability. *Teaching and Learning in Medicine, 6*, 45-48.

Colliver, J.A., Morrison, L.J., Markwell, S.J., Verhulst, S.J., Steward, D.E., Dawson-Saunders, E., & Barrows, H.S. (1990). Three studies of the effect of multiple standardized patients on intercase reliability of five standardized-patient examinations. *Teaching and Learning in Medicine, 2*, 237-245.

Cope, R.T. (1995, April). *Cautionary observations on reliability and equating of forms in high stakes performance assessment: the problem of granularity.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

De Champlain, A., Macmillan, M., Klass, D., & Margolis, M. (1999). Assessing the impact of intra-site and inter-site checklist recording discrepancies on the reliability of scores obtained in a nationally administered standardized patient examination. *Academic Medicine, 74*, *s*52-*s*54.

De Champlain, A.F., Margolis, M.J., King, A., Klass, D.J. (1997). Standardized patients' accuracy in recording examinees' behaviors using checklists. *Academic Medicine, 72, s85-s87.*

Emerging trends in the use of standardized patients. (1998). *AAMC Contemporary Issues in Medical Education, 1,* 1-2.

Fitzpatrick, A.R., & Yen, W.M. (1999, April). *The effects of test length and sample size on the reliability and equating of performance assessments.* Paper presented at the meeting of the National Council on Measurement in Education, Montréal, QC.

Gessaroli, M.E., Swanson, D.B., & De Champlain, A.F. (1998, April). *Equating performance assessments using structural equation modeling.* Paper presented at the meeting of the American Educational Research association, San Diego, CA.

Hennings, S.S., & Hirsch, T.M. (1996, April). *A comparison of equating methods applied to performance assessments.* Paper presented at the meeting of the National Council on Measurement in Education, NewYork, NY.

Huynh, H., & Ferrara, S. (1994). A comparison of equal percentile and partial credit equatings for performance-based assessments composed of free-response items. *Journal of Educational Measurement, 31,* 125-141.

Kim, S.H., & Cohen, A. (1995). A minimum χ^2 method for equating tests under the graded response model. *Applied Psychological Measurement, 19,* 167-176.

Klass, D., De Champlain, A., Fletcher, E., King, A., & Macmillan, M. (1998). Development of a performance-based test of clinical skills for the United States Medical Licensing Examination. *Federation Bulletin, The Journal of Medical Licensure and Discipline, 85,* 177-185.

Linn, R.L., & Burton, E. Performance-based assessment: Implications of task specificity. (1994). *Educational Measurement: Issues and Practice, 13,* 5-15.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 49,* 269-272.

McDonald, R. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Reznick, R., Smee, S., Rothman, R., Chalmers, A., Swanson, D., Dufresne, L., Lacombe, G., Baumer, J., Poldre, P., Levasseur, L., Cohen, R., Mendez, J., Patey, P., Boudreau, D., & Bérard, M. (1992). An objective structured clinical examination for the licentiate: Report of the pilot project of the medical council of Canada. *Academic Medicine, 67*, 487-494.

Swanson, D.B., & Norcini, J.J. (1989). Factors influencing reproducibility of tests using standardized patients. *Teaching and Leaning in Medicine, 19*, 158-166.

Swanson, D.B., Norman, G.R., & Linn, R.L. (1995). Performance-based assessment: Lessons from the health professions. Educational Researcher, *24*, 5-11.

Sykes, R.C., Yen, W., & Ito, K. (1996, April). *Scaling polytomous items that have been scored by two raters*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

Tate, R. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement, 36, 336-346.*

Tzou, H. (1996, April). *Examining equating strategies for using simulated polytomous response data.* Paper presented at the meeting of the American Educational Research Association, New York, NY.

Van der Vleuten, C.P.M., & Swanson, D.B. (1990). Assessment of clinical skills with standardized patients: State of the Art. *Teaching and Learning in Medicine, 2*, 58-76.

Vu, N.V., Marcy, J.A., Colliver, S.J., Verhulst, S.J., Travis, T.A., & Barrows, H.S. (1992). Standardized (simulated) patients' accuracy in recording clinical performance checklist items. *Medical Education, 26*, 99-104.

Vu, N.V., & Barrows, H.S. (1994). Use of standardized patients in clinical assessments: Recent developments and measurement findings. *Educational Researcher, 23*, 23-30.

Vu, N.V., Marcy, J.A., Colliver, S.J., Verhulst, S.J., Travis, T.A., Barrows, H.S. (1992). Standardized (simulated) patients' accuracy in recording clinical performance check-list items. *Medical Education, 26*, 99-104.

Table 1

Misclassification errors for homogeneous proficiency condition by rater discrepancy level

Rater discrepancy	None (baseline)	+ 5% (moderate)	+ 10% (extreme)
Classification error			
False positive rate	1.8%	3.0%	3.9%
False negative rate	2.3%	0.7%	0.2%
Total	4.1%	3.7%	4.1%

Table 2

Misclassification errors for heterogeneous proficiency condition by rater discrepancy level

Rater discrepancy	None (baseline)	+ 5% (moderate)	+ 10% (extreme)
Classification error			
False positive rate	2.8%	6.0%	10.0%
False negative rate	10.2%	4.4%	1.5%
Total	13.0%	10.4%	11.5%

TM032490

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Assessing the Impact of Standardized Patient Variability on Examination Mastery-Level Decision Consistency

Author(s): André F. DeChamplain, Marc E Gessaroli, Lisa M. Floreck Rates

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature:	Printed Name/Position/Title: ANDRE DECHAMPLAIN SENIOR PSYCHOMETRICIAN
Organization/Address: NATIONAL BOARD OF MEDICAL EXAMINERS	Telephone: (215) 590-9565 FAX: (215) 590-9449
	E-Mail Address: ADECHAMPLAIN@MAIL.NBME.ORG Date: 3-16-01

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com